Award Number: W81XWH-11-1-0402

TITLE: Mammary Cancer and Activation of Transposable Elements

PRINCIPAL INVESTIGATOR: John R. Edwards, Ph.D.

CONTRACTING ORGANIZATION: Washington University
Saint Louis, MO 63130

REPORT DATE: September 2012

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
September 2012	Annual	1 September 2011 – 31 August 2012

4. TITLE AND SUBTITLE

Mammary Cancer and Activation of Transposable Elements

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0402

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John R. Edwards, Ph.D.

E-Mail: jedwards@dom.wustl.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Washington University
Saint Louis, MO 63130

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We have made significant methodological improvements to streamline our Methyl-MAPS protocol such that it uses less input DNA while substantially reducing processing time. We also developed a new computational algorithm that utilizes the entire methylation profile in the vicinity of each gene promoter to discover patterns of methylation changes that correlate with transcription. Methylation data is conventionally analyzed using sliding windows to identify regions of differential methylation, which are at best weakly correlated with expression of nearby genes. Since annotated promoters produce only weak correlations, applying such tools to understand gene regulation by demethylation of transposable elements would be difficult. Application of our method shows that when we consider the entire methylation profile around a gene promoter, we find strong correlations between methylation and transcription changes. We also developed an extension of our method that far outperforms current approaches in identifying genes whose methylation and expression changes are correlated. With simple modifications, this tool can be used with the data we will generate in year 2 (CAGE to annotate TSSs, RNA-seq to provide expression information, and Methyl-MAPS for the high-resolution genome-wide methylation) to directly address the central hypotheses of this proposal.

15. SUBJECT TERMS
Breast cancer, epigenetic, DNA methylation, retrotransposon, preclinical cancer, development, deep sequencing.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	11	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

This project is designed to address the subject of mammary cancer development. The purpose of the project is to investigate molecular events occurring in the preclinical stages of mammary cancer; the results may lead to insights into cancer prevention in the future. Specifically, the project investigates the intersection between genome demethylation, retrotransposon transcriptional activity, and retrotransposon-driven transcription of cellular genes. Retrotransposon promoters are well recognized to function as alternative promoters for different cellular genes, generating chimeric transcripts that may or may not function in the same way as transcripts from the regular gene promoter. Transcriptional activation of retrotransposons is strongly linked with their CpG DNA methylation, and global genomic demethylation is one of the most common molecular changes in malignancies. This project tests the hypothesis that, in preclinical stages of tumor development, genomic demethylation leads to increased transcriptional activity of retrotransposons and this, in turn, leads to transcription of otherwise silent genes, potentially setting up molecular conditions that favor cancer development. Dr. Peaston's lab developed a genetically engineered mouse model in which a specific mammary cell population is fluorescently marked upon initial exposure to an oncogene. The marked population can then be collected for integrated analysis of gene expression, promoter usage, and DNA methylation after shorter or longer exposure to the oncogene during different stages of preclinical cancer development. Our role as collaborating PI in this project is to provide support for genome-side methylation profiling and contribute to the final data analysis of the combined methylation, expression and CAGE (measure of transcription start sites) data.

Body

The relevant sections from the Statement of Work are shown in the table below with corresponding goals and results. While waiting on the shipments of DNA samples to begin processing them (see Dr. Peaston's, collaborating PI, report for more information), we have worked on methodological improvements to streamline sample preparation and began to establish an analysis pipeline that can handle the three data types that will be produced in this proposal. This pipeline will greatly facilitate final analysis and interpretation of results for this project in year 2. These advancements and their importance to the project are described below. In particular we feel confident that the streamlined protocol and analysis tools for Methyl-MAPS that we have developed will easily allow us to complete all Methyl-MAPS analyses and final computational analyses outlined in the Statement of Work before the project end.

Year 1: Items from Statement of Work Relevant to Edwards Lab.

Months	Goal		Result
1-3 4-6 7-9 10-12	4. 9. 5. 8.	• Set up schedule for formal monthly electronic lab meeting between Peaston lab and Edwards lab. And regularly hold meetings.	• An informal schedule was set up for the first year with a plan for regular meetings to commence after the first DNA shipments are sent
4-6	6.	• Preliminary Methyl-MAPS analysis of pilot virgin samples	• Awaiting initial DNA shipment, which should occur shortly.
10-12	7.	• Methyl-MAPS library preparation and sequencing for replicate #1 uniparous & triparous control and tumor-prone (likely to continue to next quarter)	• Awaiting initial DNA shipment, which should occur shortly.

Improvements to Methyl-MAPS

The improved Methyl-MAPS protocol is fundamentally the same method of methylation analysis as before, but with a few small modifications that reduce the amount of input material to less than 2.5 ug and speed the library construction process (**Fig. 1**). Methylation-sensitive and methylation-dependent endonucleases are used to completely digest the genome leaving unmethylated or methylated DNA compartments respectively. After digestion, paired-end sequencing libraries are built and paired-end sequencing is performed. Paired-end sequence tags are then computationally matched to the genome to determine regions that are methylated or not. Into our protocol we have incorporated the latest developments in paired-end library construction from Life-Technologies Corp (User Guide for Mate-Paired Library Preparation on 5500 Series SOLiD Systems).

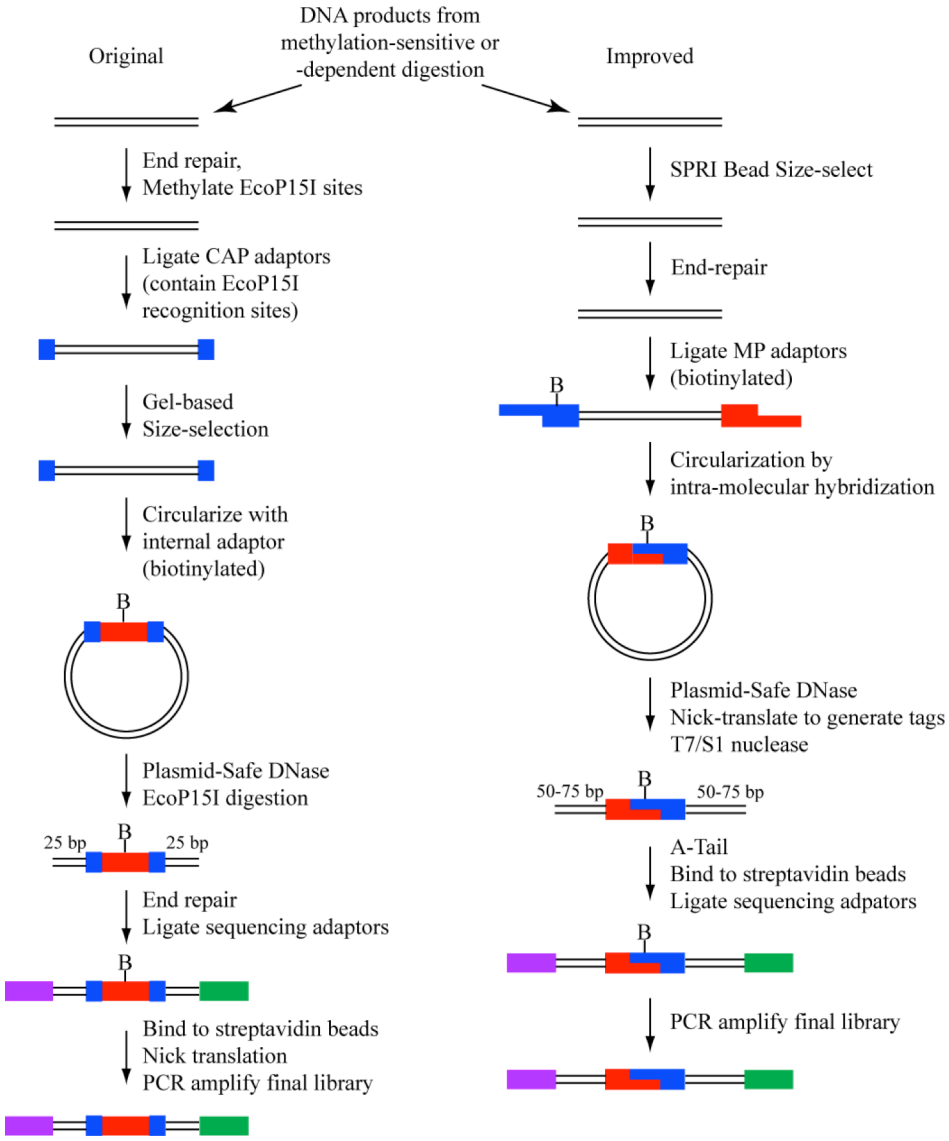

Figure 1 Comparison of the original and improved Methyl-MAPS library construction protocols. The original protocol uses EcoP15I, a type III restriction endonuclease that cleaves outwardly 25 bp to generate sequencing tags. The new protocol uses a much more efficient nick-translation tag generation step that leaves variably length tags, but whose tags are within the read lengths of current DNA sequencers (HiSeq and SOLiD). Other improvements are described in the text. The final assembled internal adaptor (red and blue) has the identical sequence in both protocols. B = Biotin.

Here we briefly describe the modifications made in the improved protocol. After enzymatic digestions we use SPRI beads to clean up the reactions. Testing revealed this method to take substantially less time then the standard method of phenol-cloroform extractions followed by ethanol precipitations with less loss of material. We subsequently use the SPRI beads to size select fragments only >600 bp. After size selection, fragments are end-repaired and special adaptors are ligated to the ends. The solution is diluted out and heated to allow the sticky ends of the adapters to come together and self-hybridize thus circularizing each molecule, while leaving a nick on each side of the assembled adaptor. Tags are produced by a timed nick translation reaction at 5° C followed by T7/S1 nuclease digestion. Library fragments are A-tailed and sequencing adapters attached. The adapter-ligated fragments are captured on streptavidin beads, purified and the final library is PCR amplified. The major changes were to three highly inefficient and time-consuming steps from the original version of the protocol. The first is the change from a gel-based size selection to a bead-based size selection, the second is the change from an *intra*-molecular circularization to an *inter*-molecular circularization and the third was a change from creating

sequence tags using the type III restriction enzyme EcoP15I, which was highly inefficient, to the very efficient nick translation tag generation.

We initially tested this new protocol with DNA from the breast cancer cell line T47D which we had analyzed using our old protocol. The initial results showed that we could reduce our input DNA requirement to 2.5 ug, without a decrease in library complexity (i.e. the number of unique DNA molecules in the library, as opposed to redundant molecules produced through PCR). In fact the complexity levels of this library were superior to those of any Methyl-MAPS library we have made, and suggests that we could potentially use even less material. We have already implemented a revised informatics pipeline that can handle the variably length tags. We found that the correlation coefficient between bisulfite-based Infinium bead array data for T47D cells and data from the new protocol improved substantially.

Computational Pipeline Improvements

We have also begun development of new computational tools to use RNA-seq and genome-wide methylation data to determine how hypomethylation in repeats and alternative- and tissue-specific promoters affects the transcriptome. Currently, the most common approach for characterizing methylation changes between two samples employs a sliding window to identify differentially methylated regions (DMRs)[1,2]. A gene with a hypermethylated DMR near its promoter is assumed to exhibit a decrease in expression, while a gene with a hypomethylated DMR should exhibit an increase in expression. In practice, the Pearson correlation coefficient between the methylation level of the DMR and the expression of its associated gene is typically no stronger than -0.4[3]. It has been assumed that better anticorrelation is precluded due to noise from experimental error, mixed cellular populations, copy number variations, chromatin modifiers or other regulation events. Another possibility, however, is that contemporary analysis methods are not sophisticated enough to recognize correlations involving more complex methylation patterns. In particular, the DMR method has several limitations. Like most existing approaches, it distills large regions of high-resolution methylation information into a single aggregate value. Values beyond some threshold are assumed to affect their nearby genes equivalently. When both hyper- and hypomethylated DMRs are present near a promoter, the method attempts to associate the best DMR instead of considering the full pattern of variation. Fundamentally, the DMR method assumes that the patterns of methylation we will find in new datasets will be the same as the examples known from past experiments. Since this approach produces poor correlations with annotated genes, we felt that to examine alternative and retrotransposon promoters we would need a better approach.

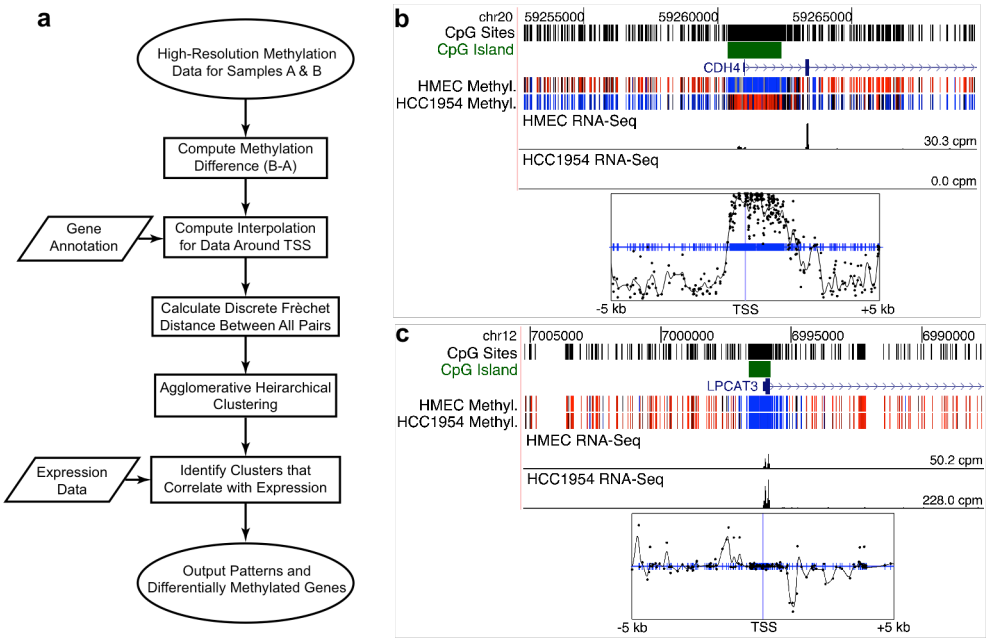

Figure 2 (**a**) Overview of the method. (**b**,**c**) Example methylation signatures from HMEC-HCC1954 BS-Seq data. Top panels show methylation and RNA-Seq expression data on the UCSC genome browser. Bottom panels show interpolated and smoothed methylation signatures (black curve) that are used to calculate the discrete Fréchet distance. Blue tick marks show locations of all CpG sites. Black dots mark experimentally measured differences in methylation between the two samples.

An outline of our new method is shown in Figure 2a. Below we demonstrate how this method works for annotated promoters, but the method can be easily adapted to the goals of this project by utilizing the CAGE data to annotate transcription start sites (TSSs), RNA-seq data to provide expression information, and Methyl-MAPS data for the high-resolution genome-wide methylation profiles.

Our method differs from other approaches by using the entire differential methylation profile in the vicinity of a gene's promoter to discover how DNA methylation changes affect gene regulation (**Fig. 2**). We represent the differential methylation for a fixed area around each gene's TSS as a continuous curve, or signature, capturing the shape of the methylation changes. We then apply a curve similarity metric, the discrete Fréchet distance, to compare differential methylation signatures for all genes. Using an unsupervised clustering technique, we arrange the signatures according to their shapes and identify which clusters of signatures are significantly correlated with gene expression changes. Generalized patterns of differential methylation can be extrapolated from the resultant clusters. Since the approach is unsupervised, there is no need for any additional assumptions about the direction of the correlations. While designed for pattern discovery, the method is easily extended to identify a list of genes potentially regulated by methylation. These gene lists are of markedly higher quality and length than those generated by existing methods.

We evaluated our technique on three datasets with high-resolution methylation and RNA-Seq expression data. The first was whole-genome bisulfite sequencing (BS-Seq) data for normal human mammary epithelial cells (HMEC) and breast cancer cells (HCC1954)[1] containing methylation levels at 84.7% of genomic CpGs with a coverage level of at least 10 in each sample. To examine how the method performed on a lower coverage dataset, we examined BS-Seq data for H1 embryonic stem cells and IMR90 fetal lung fibroblasts[2]. While the genomic coverage level of this data was high, the data was sparsely sampled at promoters: fewer than 40% of CpGs had coverage of at least 10. By including all CpGs with coverage as low as a single read, the data covered 93.5% of genomic CpGs. Expectedly, low coverage sites were extremely noisy. Lastly, we applied our method to Methyl-MAPS data from MCF7 and T47D breast cancer cells to determine its performance analyzing data generated by the method used in this grant project. We limited our analysis to sites interrogated by both digests, which included 24.9% of genomic CpGs with coverage of at least 5. Clustering results from HMEC-HCC1954 data are shown in **Figure 3**.

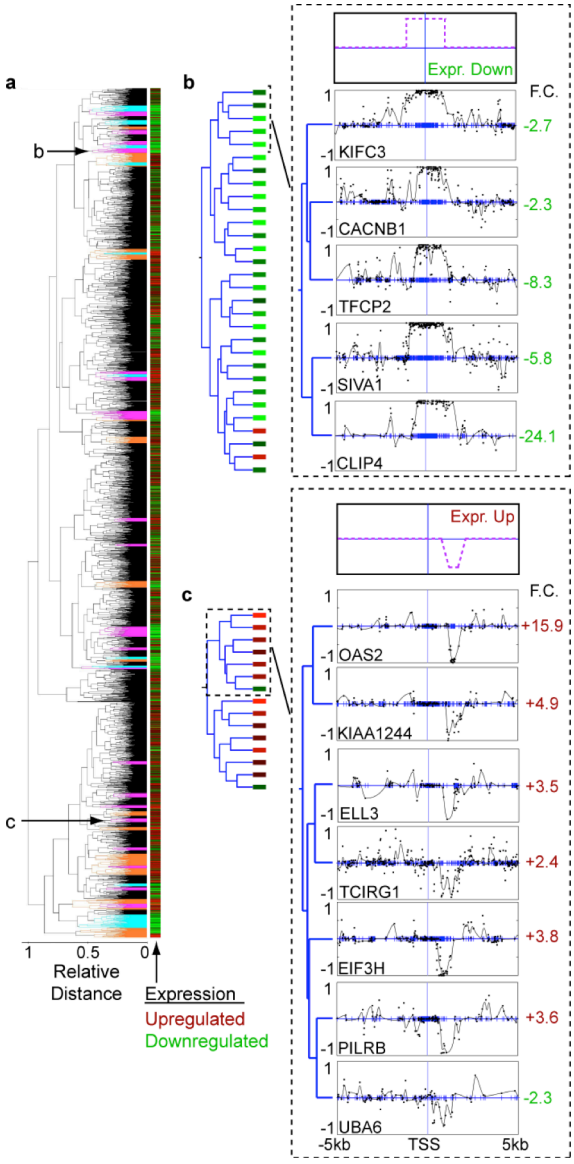

Figure 3 (**a**) Complete dendrogram for clustering 3,566 methylation signatures from HMEC-HCC1954 BS-seq data. Clusters highlighted in orange, magenta, and cyan indicate significant clusters with purity greater than 0.75, 0.85, and 0.95, respectively. Sub-clusters featured in (**b**,**c**) are indicated with arrows. A heat map of expression data is plotted in bars alongside the dendrogram. (**b**) Sub-cluster showing a pattern of methylation increase across the TSS. Left side shows the complete cluster with boxes to indicate expression. Right is one sub-cluster of patterns. (**c**) Sub-cluster showing a pattern defined by decrease in methylation 3' of the TSS. F.C. denotes expression fold change; green indicates down-regulation, red indicates up-regulation.

From the resulting sets of significant clusters, we sought to characterize the common features of the methylation signatures that may be responsible for the observed correlations with expression change. As expected, many clusters contained patterns with a strong spike of hyper- or hypomethylation across the TSS that negatively correlated with expression change (**Fig. 3b**). In addition to patterns with differential methylation across the TSS, our method identified multiple clusters in all three datasets characterized by a change in methylation downstream of the TSS (**Fig. 3c**). Meta-gene analysis of all genes in clusters associated with this pattern indicates that it operates within 3 kb of the TSS. This 3' pattern occurs across several significant clusters due to variations in other parts of the differential methylation curves. Limiting HMEC-HCC1954 BS-Seq data to only sites probed by Methyl-MAPS showed the data reduction had no impact on the ability to detect each of the identified patterns. We were also able to detect all the same pattern types in Methyl-MAPS data as were found in BS-Seq data demonstrating that in Methyl-MAPS data there is little impact on one's ability to detect patterns. As a negative control, we randomly scrambled the expression values for all genes in each dataset for 1000 random permutations. Our technique identified a false significant cluster in 1.7-2.3% of experiments (depending on the dataset) in line with the imposed statistical controls.

Gene List Generation

Enumerating a list of genes for which expression and methylation changes are potentially linked is a primary interest of any genome-wide methylation profiling experiment. Our method as outlined above is tuned to discover patterns, and thus produces a conservative gene list potentially prone to false positives. To produce a better gene list, we execute our method on a set of overlapping 5kb windows centered at a range of locations around the TSS. We identify the set of genes identified as positives (i.e. changing in the same direction as the majority of their cluster) for each window. A final list is created of all genes that are identified for at least two windows. The resulting list is more comprehensive and less prone to false positives. Judging the quality of any gene list is difficult, since there is no gold standard dataset for which the relationship between methylation and expression is well-known for all genes. To determine the extent to which genes are falsely included due to the creation of errant clusters, we randomly scrambled the expression values in the HMEC-HCC1954 dendrogram. For 1000 such experiments using default clustering parameters, only seven experiments returned any false positive genes: six reported a single false positive and a seventh returned two.

We compared the quality of the gene lists produced by our approach to lists constructed by two commonly used methods. For a DMR-based approach, regions of differential methylation are defined between two samples using a sliding window. DMRs are coupled to a particular gene using a distance cutoff [2,4]. For a promoter-based approach, a fixed window around each gene's TSS defines the gene's promoter. If methylation changes substantially within this window, the gene is labeled as differentially methylated[5,6]. We optimized DMR- and promoter-based approaches for each dataset using 69,360 and 6,174 parameter choices, respectively, while using a single common set of parameters for our approach across all datasets. Since no experimental dataset exists for which the effect of methylation change on expression is known for every gene, we evaluated each list by examining the tradeoff between the total number of differentially expressed genes identified as potentially correlated versus the fraction of identified genes that are actually correlated in the predicted direction (**Fig. 4**). This tradeoff is somewhat analogous to comparing the rate of positives to the rate of false positives, while the false negative rate is necessarily unknown. On the basis of these criteria our approach clearly dominates over the DMR- and promoter-based methods. For instance, if you set the algorithm parameters such that 20% of differentially expressed genes have a correlated differential methylation event, then for HMEC-HCC1954 data by the DMR approach only 68% of the time will the methylation and expression be anti-correlated as expected. However, using our approach, at the same 20% level, 93% of the time will silencing patterns be associated with a decrease in expression and vice-a-versa.

The optimal parameter choices for DMR- and promoter-based approaches varied widely between datasets and must be re-calibrated for each experiment. Such optimization is not necessary with our method, which performs well for a fixed set of parameters across all datasets. While our approach performs similarly well in low-

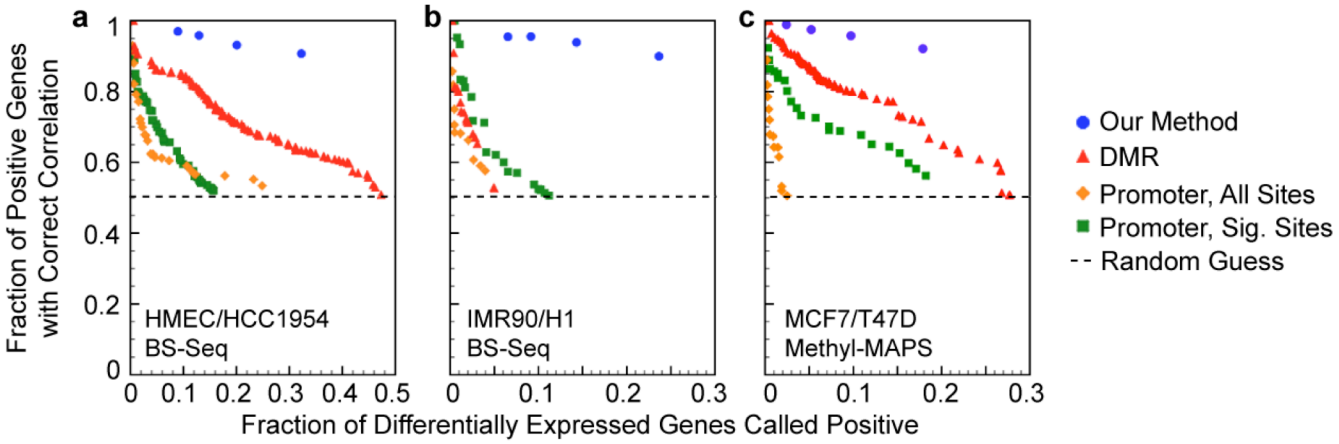

Figure 4 Comparison of gene lists generated using our approach with those from optimized DMR and promoter-centric methods for (**a**) HMEC-HCC1954 BS-Seq, (**b**) IMR90-H1 BS-Seq, and (**c**) MCF7-T47D Methyl-MAPS data. The plot shows the trade-off between the number of genes predicted to have differential expression based on their methylation (x-axis) and the quality of the predictions (y-axis). Points up and to the right indicate better performance; 50% quality is equivalent to random guessing. Only optimal parameter choices with an inverse correlation between methylation and expression are shown for DMR- and promoter-based approaches. Promoter-based approaches were optimized across both all CpG sites (All Sites) and all significant CpG sites (All Sig. Sites). A single, default set of parameters was used for our method across all three datasets.

coverage datasets, DMR- and promoter-based methods struggle (**Fig. 4b**). By downsampling the HMEC-HCC1954 dataset, we determined that BS-Seq data obtained with an average coverage as low as 7 results in little loss in our method's ability to identify genes. Methylation scores for 50% of CpGs can be removed from the data and 94.5% of genes can still be detected. These results indicate that our technique is quite robust even when much data is missing.

Our findings suggest that the role of DNA methylation cannot be fully described by simply characterizing every gene as "methylated" or "unmethylated". Using our new method, we have found and described a variety of methylation patterns that correlate with expression change. The true power of this method is in its ability to discover and separate distinct patterns without *a priori* knowledge about existing correlations, which cannot be accomplished with contemporary approaches. This allows us to realize the full potential of unbiased genome-wide profiling of DNA methylation to reveal previously unknown information about methylation's functional role. This ability will be especially important when examining regulatory elements within retroelements as will be performed in this work.

One additional implication of these results also becomes clear. The simplified models used in prior approaches at best produce weak correlations. However, if one considers a more formal description of the underlying patterns of methylation changes, methylation and expression data are highly correlated. An initial manuscript on these findings has been submitted. This tool was designed to start from a list of expression data, corresponding transcription start sites (TSSs) and high-resolution genome-wide methylation data such as from Methyl-MAPS. Thus it fits perfectly into the framework of this proposal where we will have RNA-seq data for expression, CAGE data to mark the TSSs and Methyl-MAPS methylation data for each sample. The adaptations to accommodate these datasets to address the regulation of retrotransposons are straightforward and we will have an established pipeline in place and ready for the data as it is produced in year 2.

Key Research Accomplishments

- Streamlined Methyl-MAPS protocol to use less input DNA and take less time to process samples
- Developed new computational tool to combine genome-wide expression and methylation data to output a list of genes where methylation likely contributes to their silencing or activation.

Reportable Outcomes

Manuscripts

VanderKraats ND, Hiken JF, Decker KF, Edwards JR. (2012) "Discovery of DNA methylation patterns that strongly correlate with expression changes in genome-wide high-resolution methylation data." *Submitted.*

Abstracts

VanderKraats ND, Hiken JF, Decker KF, Edwards JR. (2012) "Characterization of DNA methylation patterns that predict expression changes in genome-wide high-resolution methylation data." *Epigenetics & Chromatin Meeting at Cold Spring Harbor Laboratory*, Cold Spring Harbor, NY. Sept. 11-15.

Conclusions

The streamlined Methyl-MAPS protocol and the computational analysis pipeline we have now established will be invaluable for pushing the project ahead. The primary task for my lab was to provide Methyl-MAPS genome-wide methylation profiling for tumor DNA from the uniparous and triparous female mice from Dr. Peaston's mouse model. The tools we have in place will make this easy to accomplish in the upcoming year. The computational tools we have developed are designed to work with annotated genes as we have outlined, but can also be expanded to any transcriptional unit with a known TSS and known expression value. We will thus be able to integrate each of the datasets generated in this project (CAGE, RNA-seq and Methyl-MAPS) to address the hypothesis that, in preclinical stages of tumor development, genomic demethylation leads to increased transcriptional activity of retrotransposons and this, in turn, leads to transcription of otherwise silent genes, potentially setting up molecular conditions that favor cancer development.

References

1 Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Research* **22**, 246-258, doi:10.1101/gr.125872.111 (2012).

2 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322, doi:nature08514 [pii]10.1038/nature08514 (2009).

3 Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics* **44**, 40-46, doi:10.1038/ng.969 (2012).

4 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature genetics* **43**, 768-775, doi:10.1038/ng.865 (2011).

5 Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Research* **20**, 972-980, doi:10.1101/gr.101535.109 (2010).

6 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770, doi:nature07107 [pii]10.1038/nature07107 (2008).

Appendices

None.

Blank Page